Anthropological Review • Vol. 81(3), 252–268 (2018)

sciendo

Anthropological Review
Available online at: https://content.sciendo.com/anre
Journal homepage: www.ptantropologiczne.pl

A glance of genetic relations in the Balkan populations utilizing network analysis based on *in silico* assigned Y-DNA haplogroups

Emir Šehović[1], Martin Zieger[3], Lemana Spahić[1], Damir Marjanović[1,2], Serkan Dogan[1]*

[1]International Burch University, Department of Genetics and Bioengineering, Sarajevo, Bosnia and Herzegovina
[2]Institute for Anthropological Research, Zagreb, Croatia
[3]Institute of Forensic Medicine, Forensic Molecular Biology Dpt., University of Bern, Sulgenauweg 40, 3007 Bern, Switzerland

Abstract: The aim of this study is to provide an insight into Balkan populations' genetic relations utilizing *in silico* analysis of Y-STR haplotypes and performing haplogroup predictions together with network analysis of the same haplotypes for visualization of the relations between chosen haplotypes and Balkan populations in general. The population dataset used in this study was obtained using 23, 17, 12, 9 and 7 Y-STR loci for 13 populations. The 13 populations include: Bosnia and Herzegovina (B&H), Croatia, Macedonia, Slovenia, Greece, Romany (Hungary), Hungary, Serbia, Montenegro, Albania, Kosovo, Romania and Bulgaria. The overall dataset contains a total of 2179 samples with 1878 different haplotypes. I2a was detected as the major haplogroup in four out of thirteen analysed Balkan populations. The four populations (B&H, Croatia, Montenegro and Serbia) which had I2a as the most prevalent haplogroup were all from the former Yugoslavian republic. The remaining two major populations from former Yugoslavia, Macedonia and Slovenia, had E1b1b and R1a haplogroups as the most prevalent, respectively. The populations with E1b1b haplogroup as the most prevalent one are Macedonian, Romanian, as well as Albanian populations from Kosovo and Albania. The I2a haplogroup cluster is more compact when compared to E1b1b and R1b haplogroup clusters, indicating a larger degree of homogeneity within the haplotypes that belong to the I2a haplogroup. Our study demonstrates that a combination of haplogroup prediction and network analysis represents an effective approach to utilize publicly available Y-STR datasets for population genetics.

Key words: Balkan populations, Y-STR haplotype analysis, Haplogroup prediction, Median-joining tree, Y chromosomal haplogroups.

Introduction

The paternally inherited Y chromosome is used in forensics for identification, anthropology and population genetics for understanding origin and migrations of humans (Kayser et al. 2005 and Shi et al. 2005). The Y chromosome, despite being the smallest chromosome in the organism, still possesses two highly useful types of genetic markers, single nucleotide polymorphisms (SNPs) and short tandem repeats (STRs) (Gusmao et al. 2005;

Original Research Article Received April 10, 2018 Revised August 16, 2018 Accepted August 17, 2018
DOI: 10.2478/anre-2018-0021

Ballantyne et al. 2010; Wang et al. 2014). World population linkages, and phylogenetic trees are created based upon SNP markers, mainly because of their low mutation rate (Wang et al. 2010). The lineages determined by SNP patterns are referred to as haplogroups. Haplogroups can also be inferred from readily available Y-STR genotyping data (Athey 2006). In the forensic context there is plenty of Y-STR data available (Willuweit and Roewer 2015) that can also be explored for population genetics.

Haplogroup I referred to as "Palaeolithic" European-specific is a biological proof that Balkan population has had an additional expansion after the last Ice Age (Marjanović et al. 2005). Anatolian agriculture began to spread 8,000 to 9,500 years ago and went across the Balkan regions (Lemmen et al. 2011). Roostalu et al. (2006) and Pala et al. (2012) discussed that ancient DNA samples, especially those of mtDNA link European population to their neighbours from the Near East. Scientific efforts focused on Genetics confirmed the hypothesis that there were several waves in which farmers from Near East came to Europe due to climate changes and new farming inventions. Therefore, the overall gene pool of Europe was turbulently mixed many times. (Özdoğan, 2011; Davidović et al. 2015; Šarac et al. 2016; Veldhuis and Underdown 2017).

The Balto-Slavic speakers make approximately one third of the total Europeans and the analysis of their mitochondrial DNA and non-recombining region of the Y chromosome suggests that their genetic structure does not differ significantly from the neighbouring populations. Balto-Slavic population can be divided into East Slavs, West Slavs, and South Slavs, the latter including most of the Balkan populations (Kushniarevich et al. 2015).

Earlier research done by Šarac et al. (2018) shows that the Balkans have been a turbulent region for hundreds of years, and this region consequently contains several major haplogroups. The four major Y chromosomal haplogroups in the Balkans are I2a, E1b1b, R1a and R1b. The I2a and E1b1b haplogroups are the most abundant but the R1a and R1b still have a considerable percentage (Semino et al. 2000). Haplogroup I is thought to have appeared on the Balkan Peninsula about 45,000 years ago. However, this haplogroup most probably originated from the Middle East coming to Europe through Anatolia (Battaglia et al. 2009; Primorac et al. 2011). Haplogroup E1b1b is believed to have originated from the African continent and provides evidence of the last direct migration from Africa to Europe. (Battaglia et al. 2009). Furthermore, the inclusion of a significant percentage of R1a and R1b haplogroups within the Balkan populations confirm its historical intertwining with the other populations of Europe (Semino et al. 2000).

The aim of this study is to provide an insight into Balkan populations' genetic relations utilizing in silico analysis of Y-STR haplotypes by performing haplogroup predictions and network analysis of the same haplotypes for visualizing the relations between the chosen haplotypes and Balkan populations in general. Hence, the in silico analysis, while also being affordable, is very useful in analysing the haplogroup distributions within the Balkan countries. In addition, the analysed haplotypes were also studied using median-joining trees according to various sets of loci.

Material and methods

The population dataset used in this study was obtained using 23, 17, 12, 9 and 7 Y-STR loci for 13 populations. A 23 Y-STR set was analysed in seven populations including Bosnia and Herzegovina (Purps et al. 2014), Croatia (Purps et al. 2014), Macedonia (Purps et al. 2014), Slovenia (Purps et al. 2014), Greece (Purps et al. 2014), Romany (Hungary) (Purps et al. 2014) and Hungary (Purps et al. 2014). A set of 17 Y-STRs was analysed in two populations, namely Serbia and Montenegro (Mirabal et al. 2010). 12 Y-STR analysis includes only the Albanian population (Ferriet al. 2010). Kosovo (Peričić et al. 2004) and Romania (Barbarii et al. 2003), population data were obtained using 9 Y-STR loci analysis. Finally, 7 Y-STR analysis includes only the Bulgarian population (Zaharova et al. 2001). The reason for using some population datasets with less Y-STR loci is due to there being no publicly available data with 23 Y STR for those populations.

This dataset contains a total of 2179 samples with 1878 different haplotypes. The numbers of samples per population are not consistent, varying from 53 samples from Romany (Hungary) to 404 Montenegrin samples. The list of all analysed populations, as well as the number of samples and different haplotypes is shown in Table 1. Full data used in this article are available from the corresponding author on request.

Y chromosomal haplogroups can be assigned from Y-STR haplotypes using haplogroup predictors, which are very useful for analysing previous published Y-STR datasets (Dogan et al. 2016; Dogan et al. 2017; Heraclides et al. 2017; Gurkan et al. 2017) as well as validation purposes of haplogroup assignments based on Y SNP data (Petrejčíková et al. 2014 and Emmerova et al. 2017). The four haplogroup predictors that are in focus in this study are: Whit Athey haplogroup predictor (Athey 2006), Nevgen haplogroup predictor (Ćetković - Gentula and Nevski 2015), Vadim Urasin's

Table 1. Population datasets used in the study.

Population	Number of samples	Number of different haplotypes
Albania	339	233
Bosnia and Herzegovinian	100	100
Bulgarian	126	88
Croatian	239	239
Greek	214	214
Hungarian	100	100
Albanian (Kosovo)	117	60
Macedonian	101	101
Montenegrin	404	318
Romanian	104	97
Romany (Hungary)	53	53
Serbian	179	171
Slovenian	104	104
Total	2179	1878

haplogroup predictor (Urasin 2013) and Jim Cullen haplogroup predictor (Cullen 2008).

Overall concordance between the four haplogroup predictors as well as the concordance between each of the haplogroup predictors was analysed in order to obtain the most reliable results and to understand how the haplogroup predictors compare to each other. Relative concordance between each of the haplogroup predictors was calculated by giving a value of either 1 or 0 depending on whether the output of the predicted haplogroup is the same. Finally, dividing the obtained sum of the values by the number of haplotypes analysed will provide a relative concordance between the respective haplogroup predictors.

The haplogroup distribution percentages of all populations were obtained by calculating the average of the haplogroup percentages provided by each of the four haplogroup predictors. The haplotypes which were problematic for the haplogroup predictor to resolve were removed from the analysis.

R_{ST} pairwise matrix, as described by Slatkin (1995), was calculated on the 13 populations analysed in this study. Y-STR haplotypes are assumed to mainly follow a stepwise mutation model. Therefore, R_{ST} analysis is the most suitable in this case (Slatkin 1995). The YHRD software (yhrd.org/amova) was used in order to calculate the R_{ST} pairwise matrix as well as to create the Multidimensional Scaling (MDS) plot based on the obtained R_{ST} values (Willuweit and Roewer 2015). A minimal set of loci (DYS389I, DYS389II, DYS19, DYS391, DYS390, DYS392, DYS385, DYS393) was used in order to create the R_{ST} matrix.

The phylogenetic median-joining network algorithm is used with large sets of genetic data. It is based upon creating minimum spanning trees which are further combined into one reticulate network. In order to achieve parsimony, consensus sequences in form of median vectors, or Steiner points are added (Bendelt et al. 1999). Posada and Crandall (2001) introduced network analysis and haplogroup predictors that together make a great tool for genetic analysis of populations. Furthermore, as they complement each other's weaknesses, the results obtained from their combined overall picture can be considered more accurate than when using each method individually (Posada and Crandall 2001).

For the purpose of network analysis, a set of populations of interest were made. Each population set consisted of 15 haplotypes. The set of 15 haplotypes represented the haplogroup prediction values for each population. Hence, the percentage of haplogroups in the set of 15 haplotypes and the overall population should be nearly the same. Within specific haplogroups the haplotypes were taken arbitrarily. However, haplotypes with microvariants were not selected as they cannot be resolved by the network analysis. Furthermore, the overall concordance percentage value between all the predictors must be 100% for the haplotypes chosen; meaning that all the haplogroup predictors were concordant and predicted the same haplogroup.

Due to a small number of loci on which the Bulgarian population was analysed, it was not included within the network analysis datasets. The sets of populations included a set of all populations given in Table 1 except Bulgaria, former Yugoslavian populations and haplotype sets for individual haplogroup analyses. The network analyses including haplotypes predicted from all major haplogroups

were done by utilizing 15 haplotypes per population, while 10 haplotypes per population were used when the individual haplogroups were analysed. For the individual haplogroup network analysis new sets of haplotypes were made.

For construction of the dataset which included all populations except the Bulgarian, nine loci were used: DYS393, DYS390, DYS394, DYS385a/b, DYS439, DYS392, DYS389II, and DYS438. The locus DYS385a/b was separated into DYS385a and DYS385b. Finally, the former Yugoslavian set of populations was created using 12 loci: DYS393, DYS390, DYS394, DYS385a/b, DYS439, DYS392, DYS389II, DYS458, DYS448, DYS456, and DYS635. The reason for using more loci in the former Yugoslavian population is due to the smaller population size and a bigger likelihood of having similar or same haplotypes within the haplotype set which would yield relatively poor results. However, by using 12 loci, the haplotypes can be differentiated by a larger margin while retaining the ability of visualizing some of the similarities between them which will in the end yield proper clustering. The 12 chosen loci were selected based on the degree of variance of allele lengths they display between all the analysed populations. Loci variance ($\frac{\sum(X-\mu)^2}{N-1}$) was calculated using the VAR.S function in Excel.

Network analysis of individual haplogroups (I2a, E1b1b, R1a and R1b) of the Balkan region were analysed using the same 9 loci mentioned above. The median-joining trees were generated using the Fluxus Network 4.6 program with an equal weight on all loci and an $\epsilon = 0$ (http://www.fluxus-engineering.com/sharenet.htm) (Bandelt, Forster and Röhl 1999).

Post-processing option was included in order to calculate all the possible variations of the vectors that can be generated in the tree and select the most optimal one. In addition, all of the other versions the program has calculated can be seen. That makes it possible to analyse the median-joining trees that the program has not classified as optimal.

A modal haplotype (http://www.mymcgee.com/tools/yutility.html) was inserted in the present haplotype sets. A modal haplotype consists of the allele values with the largest number of occurrences in the haplotypes analysed. In case of a tie, the larger allele value is used.

A heatmap for each major haplogroup was created based on its respective frequency in each country (Babicki et al. 2016). Moreover, for the purpose of comparison to the median-joining trees a Principal component analysis (PCA) was performed, using the PAST program developed by Hammer (2001), on all the populations, excluding the Bulgarian population, based on the same 9 Y-STR loci used for median-joining trees.

Results

On average, the four haplogroup predictors have a 91% relative concordance between each other (haplogroup predictor concordance data not shown) as there were no Y-SNPs for referent comparison. Haplogroup assignment for each haplotype from all 4 haplogroup predictors were compared against each other and together made up the relative concordance value. The relative concordance value was calculated based only on algorithm predictions. Hence, the haplogroup percentages and results in general are considered reliable.

For the analysed populations, loci DYS385a/b, DYS481, DYS389II have a very high variance in the Balkan popula-

tions. The values correspond roughly to what would have been expected (Purps et al. 2014). It is important to emphasize that some of the loci are not covered by all populations, meaning that their statistical distribution may be skewed to a certain degree. However, the abovementioned loci cover a substantial number of haplotypes, the only exception being DYS481 with 911 haplotypes. The loci with low level of variance, indicating a similarity between the populations, are DYS391, DYS389I and GATA-H4. The loci variance analysis was useful in choosing which loci to use in the network analysis. A complete list of loci and their respective variance is shown in Table 2.

Among all the analysed populations, significant differences among the variance values between populations is found within the DYS393 locus (data for individual population variance scores for each locus not shown). The former Yugoslavian populations have a relatively low variance score with an average of 0.2602. On the other hand, other populations average around 0.5 variance score with Romany (Hungary) having the highest variance score of 1.17.

Within the DYS 438 locus, B&H, Serbian and Montenegrin population show

Table 2. Variance values of all analysed loci among the Balkan populations.

Locus	Number of populations	Variance	Number of haplotypes
DYS393	13	0.610145	1878
DYS390	13	0.862400	1878
DYS19	13	1.530143	1878
DYS391	13	0.328473	1878
DYS385a	12	3.707194	1790
DYS385b	12	3.497353	1790
DYS439	10	1.034329	1633
DYS389I	13	0.378007	1878
DYS392	13	0.842962	1878
DYS389II	13	9.029229	1878
DYS458	9	1.662084	1400
DYS437	10	0.516272	1633
DYS448	9	1.12932	1400
GATH4	9	0.550598	1400
DYS456	9	1.138877	1400
DYS576	7	1.529612	911
DYS570	7	1.971725	911
DYS438	10	0.750886	1633
DYS635	9	1.209496	1400
DYS481	7	11.02054	911
DYS533	7	0.656163	911
DYS549	7	0.727928	911
DYS643	7	1.419869	911

the least variance when compared to the other populations. Among the analysed populations the Greek and Albanian population have the highest variance score within this locus.

Within the populations which were analysed on the DYS 481 locus (seven out of the thirteen), all of them show a high degree of variance with B&H and Croatian population having the highest variance score of 14,5 and 14,3 respectively.

When analysing the loci among the former Yugoslavian populations (B&H, Croatian, Macedonian, Montenegrin, Serbian and Slovenian), DYS393, DYS391, DYS389I, and DYS437 have a relatively low variance. On the other hand, DYS385a/b, DYS19 and DYS570 have a relatively high variance, with DYS385a/b being significantly more polymorphic when compared to the other loci.

The R_{ST} pairwise matrix calculated on 13 populations, and visualized by the MDS plot, has shown a population grouping pattern which can roughly represent the geographical position of the analysed populations. As was expected, Albanian and Kosovan population positioned close to each other within the MDS plot. Furthermore, B&H and Serbian population have shown a very high degree of similarity within the MDS plot. The Romany population has appeared on the MDS plot far away from all the populations indicating a large degree of dissimilarity which is to be expected.

Another point which shows the rough geographical representation of the populations within the MDS plot is the grouping of Croatian, Slovenian and Hungarian populations in a very close fashion. The main unexpected result within the MDS plot is the Greek population grouping together with the Romanian population and being far away from the Bulgarian and Macedonian populations despite the geographical proximity.

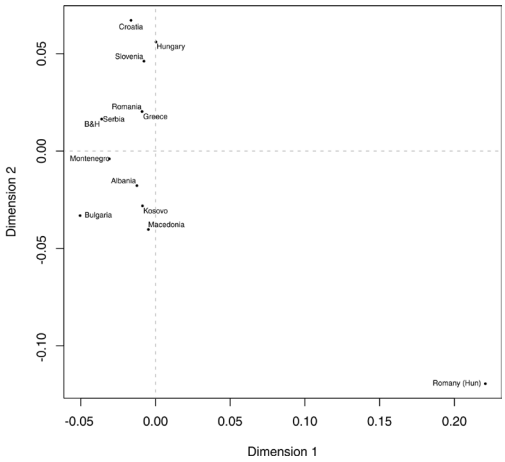

Figure 1: Two-dimensional plot of MDS analysis of R_{ST} values for Y-STR haplotypes within the 13 Balkan populations.

Haplogroup distribution of Balkan populations

I2a is detected as the major haplogroup in four out of thirteen analysed populations. Out of the six former Yugoslavian populations, four of them (B&H, Croatia, Montenegro and Serbia) have I2a as the most prevalent haplogroup, as shown in Figure 1. The other two major populations from the former Yugoslavia, Macedonia and Slovenia, have E1b1b and R1a haplogroups as the most prevalent, respectively. The second most common haplogroup for both Macedonia and Slovenia is the I2a haplogroup, confirming a degree of similarity to the other former Yugoslavian major populations.

The populations with E1b1b haplogroup as the most prevalent one are Macedonian, Romanian, as well as Albanian populations from Kosovo and Albania. Kosovo and Albania populations have a high degree of similarity in the haplogroup distribution.

I2a and E1b1b haplogroups each have a hotspot presented on the geographi-

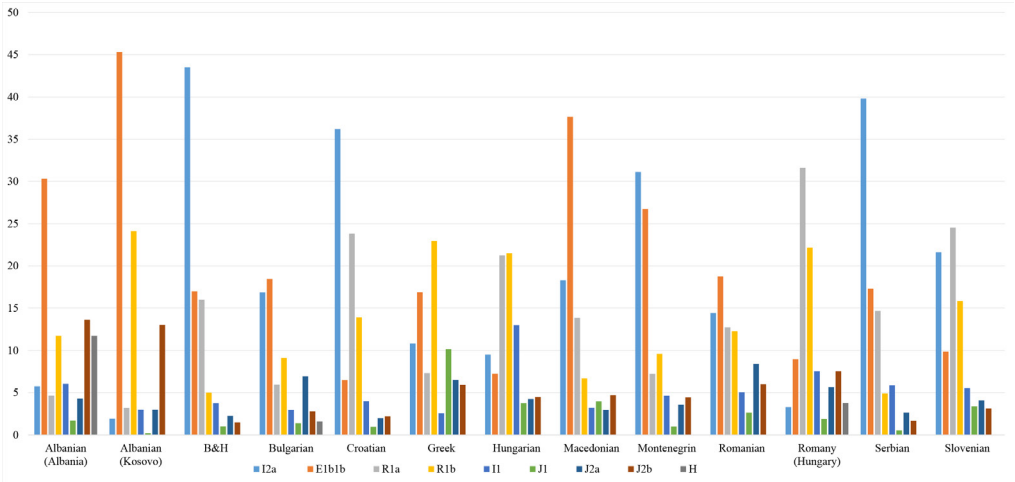

Figure 2: Distribution of haplogroups within the analysed populations.

cal heatmap shown in Figure 2. Within the R1a haplogroup a descending northwest to southeast gradient is visualized, while among the R1b haplogroup, relative to the countries in the middle Balkans, increased percentages of the R1b haplogroup can be seen in the north and south of the Balkans.

Network analysis of Balkan populations

Four major and three minor clusters are formed with the Romany (Hungarian) haplotypes forming minor independent clusters. The four major haplogroups correspond to I2a, E1b1b, R1a and R1b, while the three minor clusters are each formed by haplotypes assigned to be in the haplogroups H, J2b and I1.

As can be seen in Figure 3, the I2a haplogroup cluster shows the highest degree of relative compactness out of the analyzed haplogroups indicating a high degree of intrahaplogroup similarity between the haplotypes. On the other hand, the E1b1b haplogroup cluster has the least degree of compactness as it forms two minor clusters within it. This indicates that there is a low level of intrahaplogroup similarity between the haplotypes.

Analysis of individual major haplogroups of the Balkan region

A large cluster of B&H, Slovenia, Montenegro and Albania can be seen within the I2a haplogroup network analysis tree. Overall, the Balkan haplotypes predicted to be I2a haplogroup are the most similar to each other when compared to the other individual haplogroup trees what is the main reason for obtaining a very compact median-joining tree.

The overall E1b1b haplogroup clustering is not as compact as the I2a median-joining tree and the Balkan population haplotypes make several minor clusters (analysed minor clusters shown within the red circles) separated from the major cluster. The Romanian haplotypes are clustered with former Yugoslavian haplotypes, with E1b1b predicted haplotypes from the Romanian population tending to be the most similar to the former Yugoslavian haplotypes

The R1a haplogroup median joining tree has several major clusters found main-

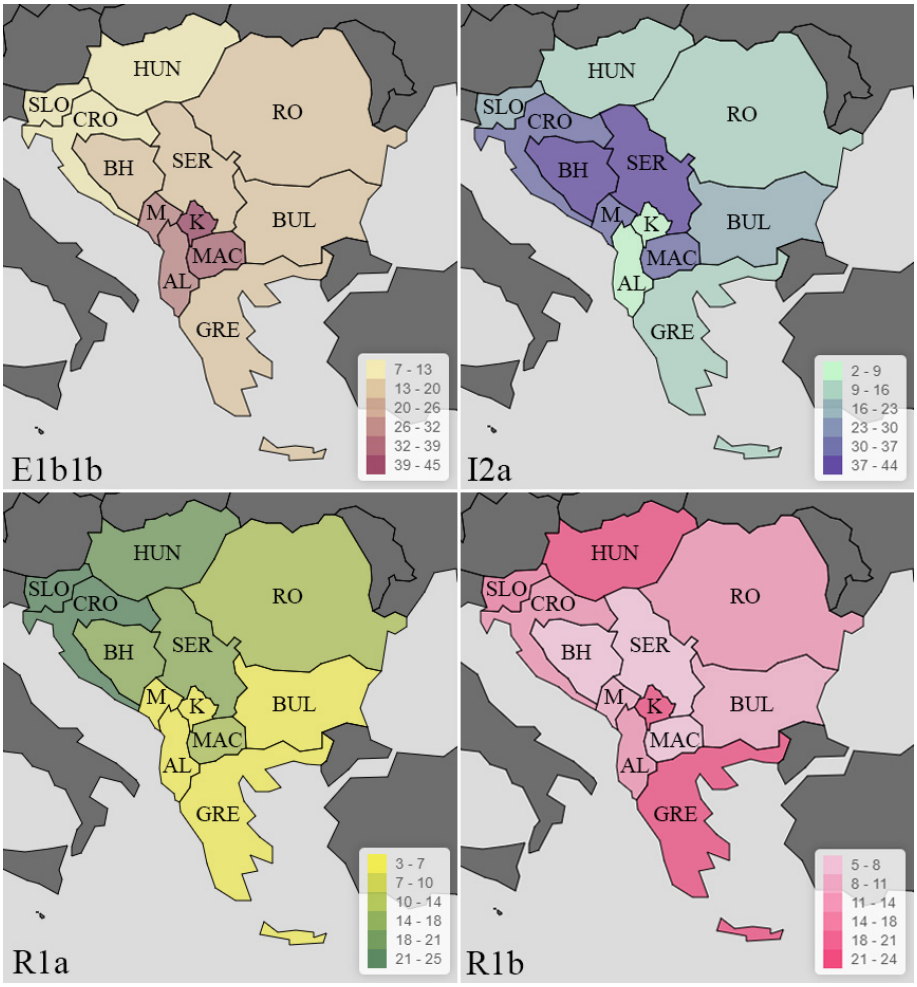

Figure 3: Geographical heatmap of the four major haplogroups within the Balkan region.

ly around the large circles of the network and is not as compact as the I2a haplogroup median joining tree as several minor clusters can be observed. Among those, two of them are population specific: one for the Romani and the other for the Romanian population. All of the major clusters are compactly clustered, while the minor clusters are not clustered together and mainly involve single population clusters.

Out of the four major individual haplogroups analysed in this study, R1b Balkan haplogroup median joining tree is the least compact one. It contains two major clusters, with the smaller major cluster consisting of mainly Kosovan, Montenegrin and Serbian along with few Greek, B&H and Macedonian haplotypes indicating a large degree of similarity between these haplotypes, and larger major cluster not showing clustering pattern of the Balkan populations but rather gradual clustering indicating that the similarities and differences between R1b predicted

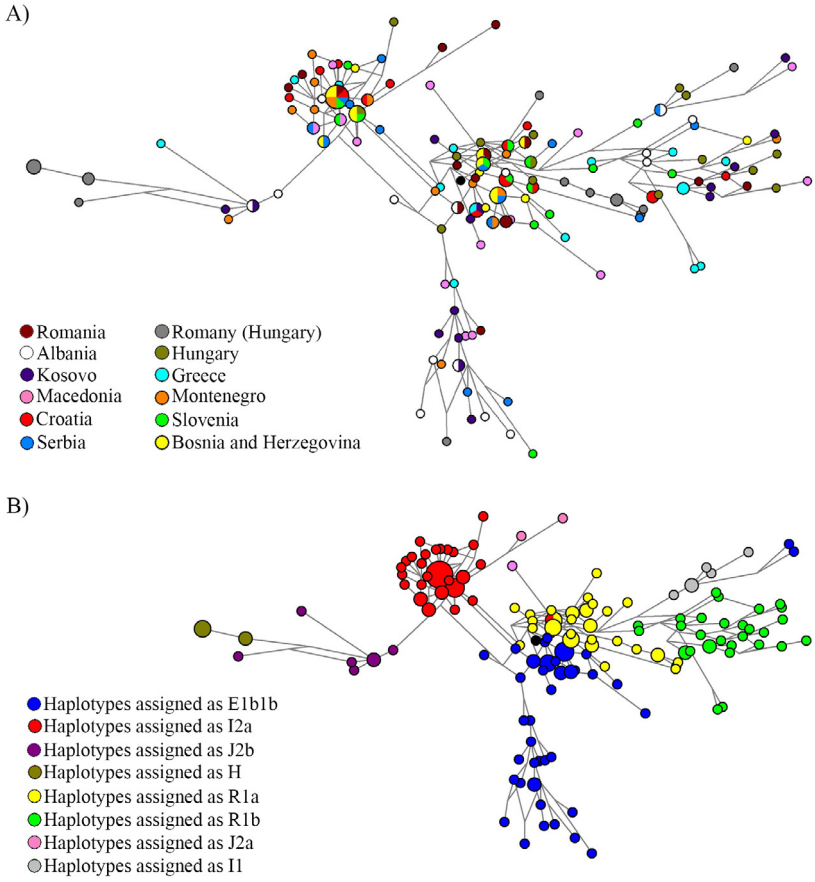

Figure 4: Median-joining network of all study populations except for Bulgaria. (A) The median-joining network showing population differentiation. (B) The median-joining network based on the assigned haplogroups.

haplotypes from the Balkan populations are of the similar extent.

PCA

Similarly to the Network analysis, the PCA largely shows haplogroup based clustering. The median-joining tree created a clearer overall picture, whereas the PCA is very useful in identifying and analysing the population distribution, minor intra-population similarities and outliers. A cluster of the Romany haplotypes, within the R1a haplogroup cluster, can be seen on the left side of Figure 6. The same cluster was also visible in the median-joining tree albeit with less clarity. Other minor population clusters that are better visually represented in the PCA are the Macedonian cluster in the bottom right (E1b1b cluster), Greece cluster on the bottom left (R1b cluster) as well as the Serbian cluster within the I2a cluster on the top of Figure 6.

The loci which contributed the most to differentiation among populations and haplogroup based clustering within the PCA are: DYS 385a/b and DYS19. The

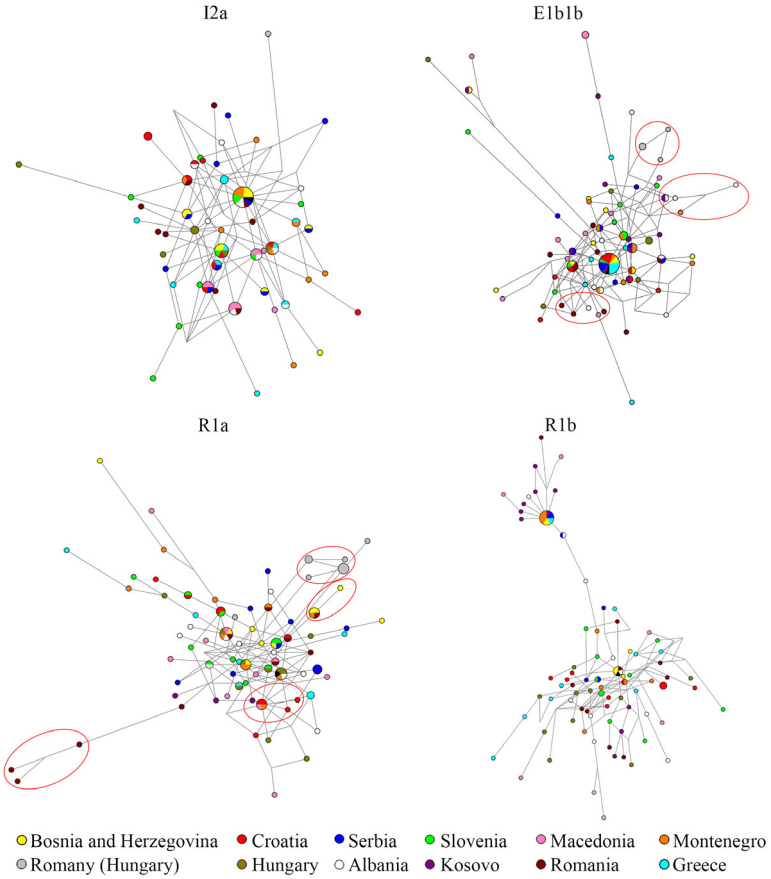

Figure 5: Median-joining network of individual major haplogroups within the Balkan region. Minor clusters marked within red circles.

respective values of loci loadings can be seen in Figure 7. Furthermore, the allele which contributed the most to the first component of the PCA was allele variant 11 on locus DYS392 while allele variant 14 on locus DYS385a contributed the most to the second component.

Discussion

The haplogroup prediction percentages of the population dataset from the current study are in expected concordance with the work of Battaglia et al. (2009). The order of major haplogroups is the same in all populations with small differences in the exact percentage values. The exact haplogroup distribution values can be seen in Table 3. This shows that the approach of utilizing four haplogroup predictors to yield accurate and reliable prediction results will enable a proper comparison and analysis of the dataset of interest.

Emmerova et al. (2017) used 5 different predictors and based on 12 STRs 3 of the predictors assigned the haplogroups with 98% accuracy compared to SNP

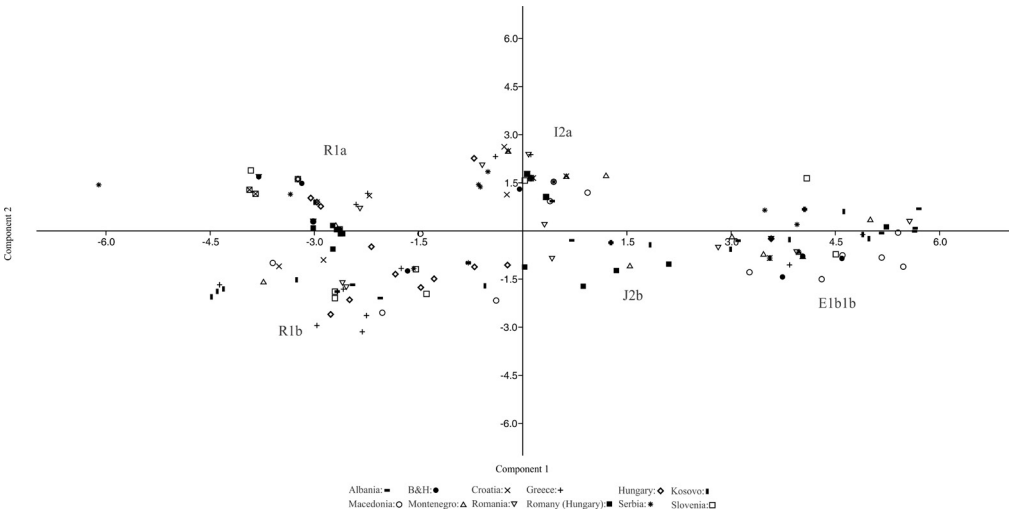

Figure 6: Principal component analysis of all the populations excluding Bulgaria based on the same 9 loci used in the network analysis.

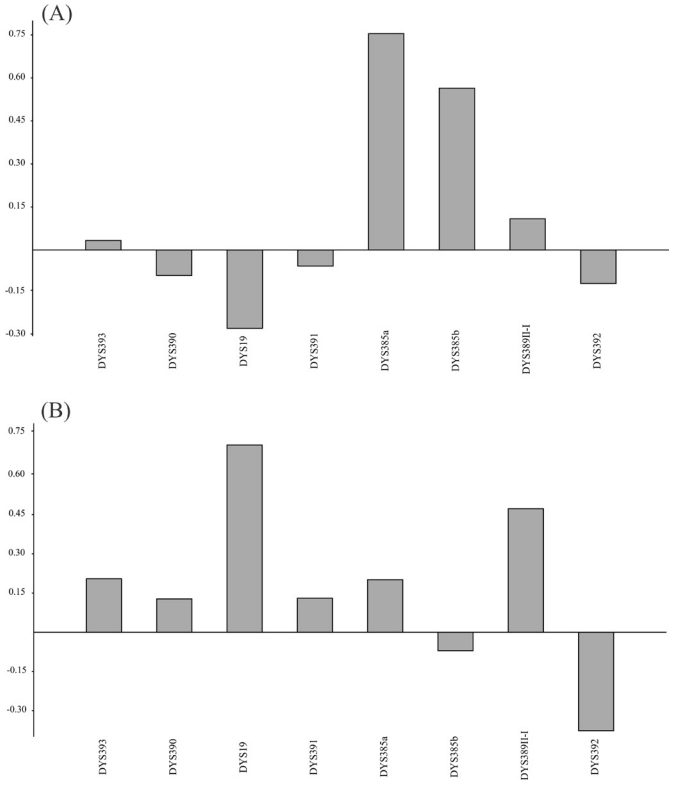

Figure 7: (A) Loci loadings of the first principal component. (B) Loci loadings of the second principal component.

typing. Petrejčíková et al. (2014) used 3 different predictors for the research based on 12 STRs and obtained 98.80%, 97.59% and 98.19% accuracy for Whit Athey´s haplogroup predictor, Jim Cullen´s haplogroup predictor and Vadim Urasin´s haplogroup predictor, respectively. Furthermore, Dogan et al. (2016) used 4 haplogroup predictors for analysing the B&H population and obtained a 99% average accuracy while the current study obtained a 91% average accuracy for all populations. This discrepancy can be attributed to the fact that many of the populations were analysed on fewer than 12 loci.

An interesting detail of our study is that compared to the 22% of haplogroup H we found in the Romany living in Hungary, Romany living in Macedonia have been shown to have 60% of the haplogroup H (Peričić et al. 2005) and only 2% of R1a and R1b haplogroups respectively.

Discrepancies observed within the I haplogroup distributions between the data used in the present study compared to Battaglia et al. (2009) for the Bosnian and Croatian populations and Regueiro et al. (2012) for the Serbian population can mainly be attributed to inaccuracies of haplogroup predictors in differentiating between the I haplogroup subclades (I1, I2a and I2b in this case).

The only notable difference to the published haplogroup distributions from Battaglia et al. (2009) was in Hungarian R1a haplogroup percentage, since the prediction algorithms approximate this haplogroup frequency to be 21% of all Hungarian Y chromosomes for the data from Purps et al. (2014), while Battaglia et al. (2009) state that it is 56.6%. A large discrepancy, that could potentially be attributed to a sampling error where one or the other population sample might not be geographically representative or might include a significant number of immigrants. The other populations that were also covered in the same article demonstrate only minor differences in few percentages which

Table 3. Haplogroup distribution within the analyzed populations. Values shown in percentages.

Population	I2a	E1b1b	R1a	R1b	I1	J1	J2a	J2b	H
Alb (Alb)	5.7	30.3	4.6	11.7	6	1.6	4.2	13.6	11.7
Alb (Kos)	1.9	45.2	3.2	24.1	2.9	0.2	2.9	13	
B&H	43.5	17	16	5	3.7	1	2.2	1.5	
Bulgarian	16.8	18.4	5.9	9.1	2.9	1.3	6.9	2.7	1.5
Croatian	36.1	6.4	23.8	13.9	3.9	0.9	1.9	2.1	
Greeek	10.8	16.8	7.3	22.9	2.5	10.1	6.5	5.9	
Hungarian	9.5	7.2	21.2	21.5	13	3.7	4.2	4.5	
Macedonian	18.3	37.6	13.8	6.6	3.2	3.9	2.9	4.7	
Montenegrin	31.1	26.7	7.2	9.5	4.6	0.9	3.5	4.4	
Romanian	14.4	18.7	12.7	12.2	5	2.6	8.4	6	
Romany (Hu)	3.3	8.9	31.6	22.1	7.5	1.8	5.6	7.5	3.7
Serbian	39.8	17.3	14.6	4.8	5.8	0.5	2.6	1.6	
Slovenian	21.6	9.8	24.5	15.8	5.5	3.3	4	3.1	

are justifiable and can be attributed to the small sample size of this population possibly including a number of rare haplotypes which are not easily resolved by the predictors.

The results generated from the current study show that the similarities between the former Yugoslavian countries are, as expected, present in a larger extent than between the other countries. The only population which showed higher level of dissimilarity to the other former Yugoslavian populations is Slovenia. The major Slovenian haplogroup is, unlike the other Balkan countries, R1a.

Kosovan and Albanian populations have shown a high degree of similarity which was expected considering their common history, language and shared demographics (Belledi et al. 2000; Peričić et al. 2004). Furthermore, when the Kosovan population was removed, Albanian haplotypes were found to cluster with the Greek population (data not shown) which does have certain logical reasoning behind it since Albania and Greece share borders.

Ballantyne et al. (2014) performed a similar network analysis using 10 and 15 loci with 1000 haplotypes randomly selected from the total dataset and compared to the median joining tree used in this study where only 10-15 haplotypes per population were used. This shows that the mentioned network analysis is very detailed. However, it is challenging to visualize such a tree. Hence, we have optimized our method and used less haplotypes per median-joining tree.

Conclusion

Various historical and immigration events have shaped the demographic picture of the Balkan region, making it very diverse from the genetical perspective. Our study confirms that relative geographical haplogroup distribution is the most informative way of interpreting Y-chromosomal population data and can be reliably done in silico by utilizing large publicly available STR datasets stored in repositories such as YHRD (Willuweit and Roewer 2015). Here we demonstrate, once again, that geographical distance between the populations, as would be expected, is one of the key factors in shaping the genetic similarities or differences between them. The data we obtain with this in silico approach is in concordance with the existing literature.

The usage of four haplogroup predictors, providing very reliable and accurate major haplogroup predictions, enabled us to create highly accurate median-joining trees. In silico haplogroup assignments and network analysis can be combined to detect emerging subclusters for selective SNP analysis for the investigation of possible subclades and to specifically target uncertain haplotypes that should be ultimately confirmed by SNP analysis. This combined approach provides a very cost-effective Y haplogroup analysis, since SNP typing of the whole dataset can be avoided without a substantial loss of information. However, Y-SNP genotyping should still remain the main and standard method of choice for haplogroup determination.

Authors' contributions

EŠ was involved in conceptualization, data analysis and curation, methodology, visualization, writing original draft; MZ was involved in methodology, writing original draft; LS was involved in writing original draft; DM was involved in supervision, reviewing and editing the ma-

nuscript; SD was involved in conceptualization, methodology, data curation, visualization, and writing original draft.

Conflict of interest

The authors declare no conflict of interest.

Corresponding author
Emir Šehović. Address: Izeta Karšića 54, Sarajevo, Bosnia and Herzegovina
Email address: emir.sehovic@gmail.com.

References

Athey TW. 2006. Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach. J Genet Geneal 2:34-39.

Babicki S, Arndt D, Marcu A, Liang Y, Grant JR, Maciejewski A, Wishart DS. 2016. Heatmapper: web-enabled heat mapping for all. Nucleic Acids Res. (epub ahead of print). doi:10.1093/nar/gkw419

Ballantyne K. N, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, et al. 2010. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. The American Journal of Human Genetics 87(3): 341-353.

Ballantyne KN, Ralf A, Aboukhalid R, Achakzai NM, Anjos MJ, Ayub Q, et al.2014. Toward Male Individualization with Rapidly Mutating Y -Chromosomal Short Tandem Repeats. Human mutation 35(8):1021-1032.

Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Molecular biology and evolution 16(1):37-48.

Barbarii LE, Burkhard R, Dan Dermengiu D. Y-chromosomal STR haplotypes in a Romanian population sample. 2003. International journal of legal medicine 117(5):312-315.

Bar-Yosef Ofer, 2002. The Upper Palaeolithic Revolution. Annual Reviews Anthropology 31:1, 363-393

Battaglia V, Fornarino S, Al-Zahery N, Olivieri A, Pala M, Myres NM, et al. 2009. Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. European Journal of Human Genetics 17(6):820-830.

Belledi M, Poloni ES, Casalotti R, Conterio F, Mikerezi I, Tagliavini J, et al. 2000.Maternal and paternal lineages in Albania and the genetic structure of Indo-European populations. European journal of human genetics: EJHG 8(7):480.

Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, et al. 2012. Mapping the origins and expansion of the Indo-European language family. Science 337(6097):957-960.

Cullen J, 2008. World Haplogroup and Haplo-I Subclade Predictor. www.members.bex.net/jtcullen515/haplotest.htm. Accessed 27 May 2016.

Ćetković, Gentula M, Nevski A. 2015.Y-DNA haplogroup predictor – NevGen. www.nevgen.org/. Accessed 27 May 2016.

Davidović S, Malyarchuk B, Aleksic J. M, Derenko M, Topalovic V, et al. 2015. Mitochondrial DNA perspective of Serbian genetic diversity. American journal of physical anthropology 156(3), 449-465.

Doğan S, Ašić A, Doğan G, Besic L, Marjanović D. 2016. Y-Chromosome Haplogroups in the Bosnian-Herzegovinian Population Based on 23 Y-STR Loci. Human biology 88(3):201-9.

Doğan S, Babic N, Gurkan C, Goksu A, Marjanović D, Hadziavdic V. 2016. Y-chromosomal haplogroup distribution in the Tuzla Canton of Bosnia and Herzegovina: A concordance study using four different in silico assignment algorithms based on Y-STR data. HOMO-Journal of Comparative Human Biology 1;67(6):471-83.

Doğan S, Gurkan C, Dogan M, Balkaya HE, Tunc R, Demirdov DK, Ameen NA, Marjanović D. 2017. A glimpse at the intricate mosaic of ethnicities from Mesopotamia: Paternal lineages of the Northern Iraqi Arabs, Kurds, Syriacs, Turkmens and Yazidis. PloS one 3;12(11):e0187408.

Emmerova B, Ehlera E, Comasd D, Votrubovaa J, Vanek D. 2017. Comparison of

Y-chromosomal haplogroup predictors. Forensic Science International 6:145-147.

Ferri G, Tofanelli S, Alu M, Taglioli L, Radheshi E, Corradini B, et al. 2010. Y-STR variation in Albanian populations: implications on the match probabilities and the genetic legacy of the minority claiming an Egyptian descent. International journal of legal medicine 124(5):363-370.

Grugni V, Battaglia V, Kashani BH, Parolo S, Al-Zahery N, Achilli et al. 2012. Ancient migratory events in the Middle East: new clues from the Y-chromosome variation of modern Iranians. PloS one 7(7): e41252.

Gurkan C, Sevay H, Demirdov DK, Hossoz S, Ceker D, Teralı K, Erol AS. 2017. Turkish Cypriot paternal lineages bear an autochthonous character and closest resemblance to those from neighbouring Near Eastern populations. Annals of human biology 17;44(2):164-74.

Gusmao L, Sanchez-Diz P, Calafell F, Martin P, Alonso CA, Alvarez-Fernandez F, et al. 2005. Mutation rates at Y chromosome specific microsatellites. Human Mutation 26(6):520-528.

Hammer, Ø, Harper D.A.T, Ryan P.D. 2001. PAST: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica 4(1):9.

Heraclides A, Bashiardes E, Fernández--Domínguez E, Bertoncini S, Chimonas M, Christofi V, King J, Budowle B, Manoli P, Cariolou MA. 2017. Y-chromosomal analysis of Greek Cypriots reveals a primarily common pre-Ottoman paternal ancestry with Turkish Cypriots. PloS one 16;12(6):e0179474.

Kayser M, Lao O, Anslinger K, Augustin C, Bargel G. Edelmann J, et al. 2005. Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. Human Genetics 117(5): 428-443.

Kushniarevich A, Utevska O, Chuhryaeva M, Agdzhoyan A, Dibirova K, Uktveryte I, et al. 2015. Genetic heritage of the Balto-Slavic speaking populations: a synthesis of autosomal, mitochondrial and Y-chromosomal data. PLoS One 10(9):e0135820.

Lemmen C, Gronenborn, D, & Wirtz, K. W. 2011. A simulation of the Neolithic transition in Western Eurasia. Journal of Archaeological Science 38(12), 3459-3470.

Marjanović D, Fornarino S, Montagna S, Primorac D, Hadziselimovic R, Vidovic S, et al. 2005. The peopling of modern Bosnia -Herzegovina: Y -chromosome haplogroups in the three main ethnic groups. Annals of Human Genetics 69(6): 757-763.

Mirabal S, Varljen T, Gayden T, Regueiro M, Vujović S, Popović D, et al. 2010. Human Y -chromosome short tandem repeats: A tale of acculturation and migrations as mechanisms for the diffusion of agriculture in the Balkan Peninsula. American journal of physical anthropology 142(3):380-390.

Özdoğan M. 2011. Archaeological evidence on the westward expansion of farming communities from eastern Anatolia to the Aegean and the Balkans. Current Anthropology 52(S4), S415-S430.

Pala M, Olivieri A, Achilli A, Accetturo M, Metspalu E, et al. 2012. Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. The American journal of human genetics 90(5), 915-924.

Peričić M, Barać Lauc L, Martinović Klarić I, Janićijević B, Behluli I, Rudan P. 2004. Y chromosome haplotypes in Albanian population from Kosovo. Forensic science international 146(1):61-64.

Peričić M, Barać Lauc L, Martinović Klarić I, Rootsi S, Janićević B, et al. 2005. High-Resolution Phylogenetic Analysis of Southeastern Europe Traces Major Episodes of Paternal Gene Flow Among Slavic Populations. Molecular Biology and Evolution 22(10):1964-1975

Petrejčíková E, Čarnogurská J, Hronská D, Bernasovská J, Boroňová I, Gabriková D, et al. 2014. Y-SNP analysis versus Y-haplogroup predictor in the Slovak population. Anthropologischer Anzeiger 71(3):275-285.

Posada D, Crandall KA. 2001. Intraspecific gene genealogies: trees grafting into networks. Trends in ecology & evolution

16(1):37-45.

Primorac D, Marjanović D, Rudan P, Villems R, Underhill P. A. 2011. Croatian genetic heritage: Y-chromosome story. Croatian medical journal 52(3): 225-234.

Purps J, Siegert S, Willuweit S, Nagy M, Alves C, Salazar R, et al. 2014. A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. Forensic Science International. Genetics 12: 12-23.

Regueiro M, Rivera L, Damnjanovic T, Lukovic L, Milasin J, Herrera RJ. 2012. High levels of Paleolithic Y-chromosome lineages characterize Serbia. Gene 498(1):59-67.

Roostalu U, Kutuev I, Loogväli E. L, Metspalu E, Tambets K, et al. 2006. Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. Molecular biology and evolution 24(2), 436-448.

Rootsi S, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, et al. 2004. Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. The American Journal of Human Genetics 75(1):128-137

Semino O, Passarino G, Oefner P J., Lin Alice A., Arbuzova S, et al. 2000. The Genetic Legacy of Paleolithic Homo sapiens sapiens in Extant Europeans: A Y Chromosome Perspective. Science 290 (5494):1155-1159.

Semino O, Passarino G, Brega A, Fellous M, Santachiara-Benerecetti AS. 1996. A view of the neolithicdemic diffusion in Europe through two Y chromosome-specific markers. American journal of human genetics 59(4):964.

Shi H, Dong YL, Wen B, Xiao CJ, Underhill PA, Shen PD, et al. 2005.Y-chromosome evidence of southern origin of the East Asian–specific haplogroup O3-M122. The American Journal of Human Genetics 77(3):408-419.

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139(1), 457-462.

Stevanović M, Dobricić V, Keckarević D, Perović A, Savić-Pavićević D, Keckarević-Marković M, et al. Human Y-specific STR haplotypes in population of Serbia and Montenegro. 2007. Forensic science international 171(2):216-221.

Šarac J, Auguštin D. H, Metspalu E, Novokmet N, Missoni S, Rudan P. 2018. Maternal Genetic Profile of Serbian And Montenegrin Populations from Southeastern Europe. Genetics & Applications 1(2), 14-22.

Šarac J, Šarić T, Auguštin D. H, Novokmet N, Vekarić N, Mustać M, et al. 2016. Genetic heritage of Croatians in the Southeastern European gene pool—Y chromosome analysis of the Croatian continental and Island population. American Journal of Human Biology 28(6): 837-845.

Urasin V. 2013.Y Predictor by VadimUrasin v1.5.0. http://predictor.ydna.ru/. [Accessed 27 May 2016].

Veldhuis D, Underdown S.J. (2017) Human biology of migration, Annals of Human Biology 44:5, 393-396

Wang CC, Jin L, Li H. 2014. Natural selection on human Y chromosomes. J. Genet. Genomics 41:47-52.

Wang CC, Yan S, Li H. 2010. Surnames and the Y chromosomes. Commun. Contemp. Anthropol 4:26-33.

Willuweit S, Roewer L. 2015. The new Y chromosome haplotype reference database. Forensic Science International: Genetics 15:43-48.

Zaharova B, Andonova S, Gilissen A, Cassiman JJ, Decorte R, Kremensky I. 2001. Y-chromosomal STR haplotypes in three major population groups in Bulgaria. Forensic science international 124(2):182-186.

Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, et al. 2003. The genetic legacy of the Mongols. The American Journal of Human Genetics 72(3):717-721.